GUI FOR ANALYSIS

A PROJECT REPORT

Submitted by

DHANUSHKUMAR S G(20BCS026)
DHINAKARAN S B (20BCS029)
GOKUL M (20BCS032)
KARTHIKEYAN K (20BCS050)
KARTHIKEYAN R (20BCS051)

In partial fulfilment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

KUMARAGURU COLLEGE OF TECHNOLOGY
COIMBATORE-641 049
(An Autonomous Institution Affiliated to Anna University, Chennai)

December 2022

KUMARAGURU COLLEGE OF TECHNOLOGYCOIMBATORE 641 049

(An Autonomous Institution Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report **"GUI FOR ANALYSIS "** is the bonafide work of **"DHANUSHKUMAR S G(20BCS026), DHINAKARAN S B (20BCS029), GOKUL M (20BCS032), KARTHIKEYAN K (20BCS050), KARTHIKEYAN R (20BCS051)"** who carried out the project work under my supervision.

SIGNATURE

Dr. Devaki. P, Ph.D.,

HEAD OF THE DEPARTMENT

Department of Computer Science and Engineering,
Kumaraguru College of Technology
Coimbatore – 641 049

SIGNATURE

Mr. V. Senthil Kumar AP II/CSE,

SUPERVISOR

Department of Computer Science and Engineering,
Kumaraguru College of Technology
Coimbatore – 641 049

The candidates with university register number **20BCS026, 20BCS029, 20BCS032, 20BCS050 and 20BCS051** were examined in the Project Viva-Voice examination held on 31.12.2022.

Internal Examiner

External Examiner

DECLARATION

We affirm that the project work titled **"GUI FOR ANALYSIS"** being submitted in partial fulfilment for the award of B.E Computer Science and Engineering is the original work carried out by us. It has not formed the part of any other project work submitted for the award of any degree or diploma, either in this or any other University.

DHANUSHKUMAR S G (20BCS026)

DHINAKARAN S B (20BCS029)

GOKUL M (20BCS032)

KARTHIKEYAN K (20BCS050)

KATHIKEYAN R (20BCS051)

I certify that the declaration made above by the candidates are true.

Mr. V. Senthil Kumar,
Assistant Professor II,

Department of Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore – 641 049.

ACKNOWLEDGEMENT

We express our profound gratitude to the management of Kumaraguru College of Technology for providing as with the required infrastructure that enabled us to successfully complete the project.

We extend our gratefulness to our Principal, **Dr. Saravanan D,** for providing usthe necessary facilities to pursue the project.

We would like to acknowledge **Dr. P. Devaki,** Professor and Head, Department of Computer Science and Engineering, for her support and encouragement throughout this project.

We thank our guide **Mr. V. Senthil Kumar,** Assistant Professor II, Department of Computer Science and Engineering, for his constant and continuous effort, guidance and valuable time.

Our sincere and hearty thanks to staff members of Department of Computer Science and Engineering of Kumaraguru College of Technology for their well wishes, timely help and support rendered to us during our project. We are greatly indebted to our family, relatives and friends, without whom life would have not

been shaped to this level.

ABSTRACT

Data analytics is critical since it allows firms to improve their performance. Data cleaning is the process of identifying and removing the errors in the data warehouse. While collecting and combining data from various sources into a data warehouse, ensuring high data quality and consistency becomes a significant, often expensive, and always challenging task. Without clean and correct data, the usefulness of Data Mining and data warehousing is mitigated.

Data visualization involves presenting data in graphical or pictorial form which makes the information easy to understand. It helps to explain facts and determine courses of action. It will benefit any field of study that requires innovative ways of presenting large, complex information. The advent of computer graphics has shaped modern visualization.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABBREVIATION	DESCRIPTION
EVA	Exploratory visual analysis
JS	Java Script
ML	Machine Learning
OS	Operating System
API	Application Program Interface

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION AND BACKGROUND

Automated Data Cleaning aims towards creating a framework for cleaning raw, erroneous structured datasets like Tabular Data and unstructured data like images and audio samples to yield a clean enough data sample that can be efficiently used for further processing. Any dataset is obtained by consolidating information from numerous sources which makes it prone to errors and inconsistencies. Data that is incomplete or inaccurate is known as dirty data. When dealing with real-world data, dirty data is the norm rather than the exception. Thus the various types of anomalies occurring in data that have to be eliminated.

Cleaning the data involves the removal of duplicates, the removal or replacement of missing items, the correction of mis fielded values, the maintenance of uniform formatting, and a variety of additional duties. Having clean data will eventually boost overall productivity and provide for the greatest quality information in your decision-making, which takes time. As a result, an automatic data cleansing tool is necessary.

There has been the need for displaying massive amounts of data in a way that is easily accessible and understandable. Organizations generate data every day. As a result, the amount of data available on the Web has increased dramatically. It is difficult for users to visualize, explore, and use this enormous data. The ability to visualize data is crucial to scientific research. Today, computers can be used to process large amounts of data. Data visualization is concerned with the design,

development, and application of computer-generated graphical representation of the data. It provides effective data representation of data originating from different sources. This enables decision makers to see analytics in visual form and makes it easy for them to make sense of the data. It helps them discover patterns, comprehend information, and form an opinion.

1.2 OBJECTIVES OF THE PROJECT

The project's goal is to create a data cleaning and visualization tool to smooth data analysis process. This project primarily concentrates different data visualisation techniques.

1.3 SCOPE OF THE PROJECT

This application when completely constructed will be able to automate data cleaning and data visualisation. It will be making data analysis simpler. Inferences about the data can also be collected from this framework.

CHAPTER 2

LITERATURE REVIEW

[1] Research Data Analysis with Power BI

Vijay Krishnan, S Bharanidharan, G Krishnamoorth

Power BI has taken the world of business intelligence, data visualization and analytics by storm. Power BI is an online service that enables searching data, transforming it, visualizing it, and sharing the developed reports and dashboards with other users in the same or different department/organizations or even with the general public. As of February 2017, more than 200,000 organizations across 205 countries are using Power BI. Power BI is having a free option that has adequate features and functionality, it has become a serious contender for use as a business intelligence platform in small and medium organizations. One of the innovative features of Power BI is its Quick Insights feature (Michael Hart, 2017) that is built on a growing set of advanced analytical algorithms. After uploading a dataset to Power BI, a click of a button can be used to invoke this feature that automatically builds many reports based on its analysis of the data, without any human intervention being required. This also helps reduce human errors, in calculations and statistical techniques, which lead to un-verifiable research. Accepting even Excel spreadsheets as input, Power BI is easy to use and ripe for adoption as a platform for Research Data Analysis. In this paper, an attempt has been made to show how easily a dataset of research data

can be transformed by Power BI into a set of analytical reports and dashboards, and which can be shared with ease.

[2] Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau

Supporting exploratory visual analysis (EVA) is a central goal of visualization research, and yet our understanding of the process is arguably vague and piecemeal. We contribute a consistent definition of EVA through review of the relevant literature, and an empirical evaluation of existing assumptions regarding how analysts perform EVA using Tableau, a popular visual analysis tool. We present the results of a study where 27 Tableau users answered various analysis questions across 3 datasets. We measure task performance, identify recurring patterns across participants' analyses, and assess variance from task specificity and dataset. We find striking differences between existing assumptions and the collected data. Participants successfully completed a variety of tasks, with over 80% accuracy across focused tasks with measurably correct answers. The observed cadence of analyses is surprisingly slow compared to popular assumptions from the database community. We find significant overlap in analyses across participants, showing that EVA behaviors can be predictable. Furthermore, we find few structural differences between behavior graphs for open-ended and more focused exploration tasks.

[3] The art of Data Analysis

Muhammad Ibrahim
Govt. M A O College

After collecting the accurate and reliable data successfully by using the appropriate method from the source, the next step is how to extract the pertinent and useful information buried in the data for further manipulation and interpretation. The process of performing certain calculations and evaluation in order to extract relevant information from data is called data analysis. The data analysis may take several steps to reach certain conclusions. Simple data can be organized very easily, while the complex data requires proper processing. The word “processing” means the recasting and dealing with data making
ready for analysis. This paper focuses on process and data analysis.

[4] Towards Automated Data Cleaning Workflows

Mohammad Mahdavi
Felix Neutatz
Technische Universität Berlin
Larysa Visengeriyeva
Technische Universität Berlin
Ziawasch Abedjan
Leibniz Universität Hannover

The success of AI-based technologies depends crucially on trustful and clean data. Research in data cleaning has provided a variety of approaches to address different data quality problems. Most of them require some prior knowledge about the dataset in order to select and configure the approach correctly. We argue that for unknown datasets, it is unrealistic to know the data quality problems upfront and to formulate

all necessary quality constraints in one shot. Pragmatically, the user solves data quality problems by implementing an iterative cleaning process. This incremental approach poses the challenge of identifying the right sequence of cleaning routines and their configurations. In this paper, we highlight our work in progress towards building a cleaning workflow orchestrator that learns from cleaning tasks in the past and proposes promising cleaning workflows for a new dataset. To this end, we highlight new approaches for selecting the most promising error detection routines, aggregating their outputs, and explaining the final results.

CHAPTER 3

PROBLEM DEFINITION

The most appealing reason is that one simple graph says more than twenty pages of prose. These graphs laid foundation for the growth of Data analysis - an auspicious branch of computer science. Data analysis for business, finds its significant place in the wide umbrella of its applications.

It provides actionable insights into customer behavior along with comprehensive market analysis thereby providing a competitive edge to businesses. Data analysis involves the processes of Defining the Question, Collecting the data,

Cleaning the data, Visualizing the data, Analyzing the data, Sharing your results, Embracing failure, Summary. Currently these processes are being carried out manually, tools automating any of these processes would come in handy.

CHAPTER 4

PROPOSED SYSTEM

The graphs give the visual presentation of data. Graphs are useful in model fitting of data. At looking the graphs, one can understand the data in an easy way. Most visualization designs are to aid decision making and serve as tools that augment cognition. In designing and building a data visualization prototype, one must be guided by how the visualization will be applied.

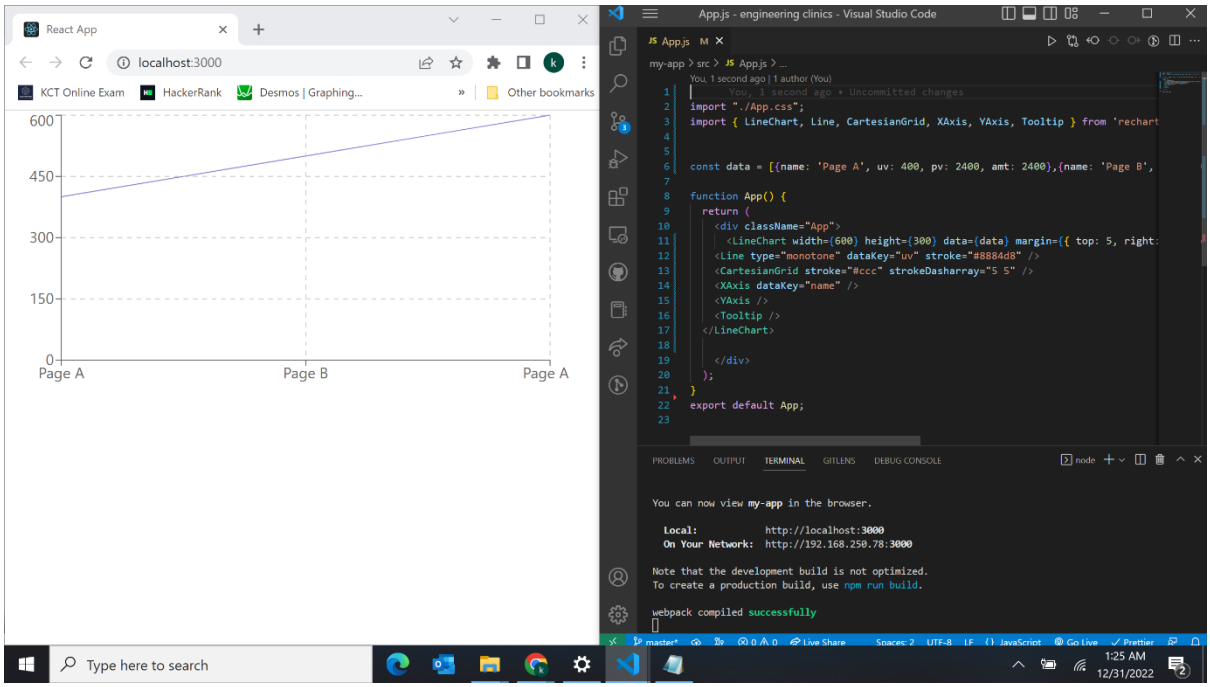
The proposed system is a framework which provide users with powerful and widely applicable tool for analysing and interpreting large and complex data.

4.1 VISUALIZATION TECHNIQUES

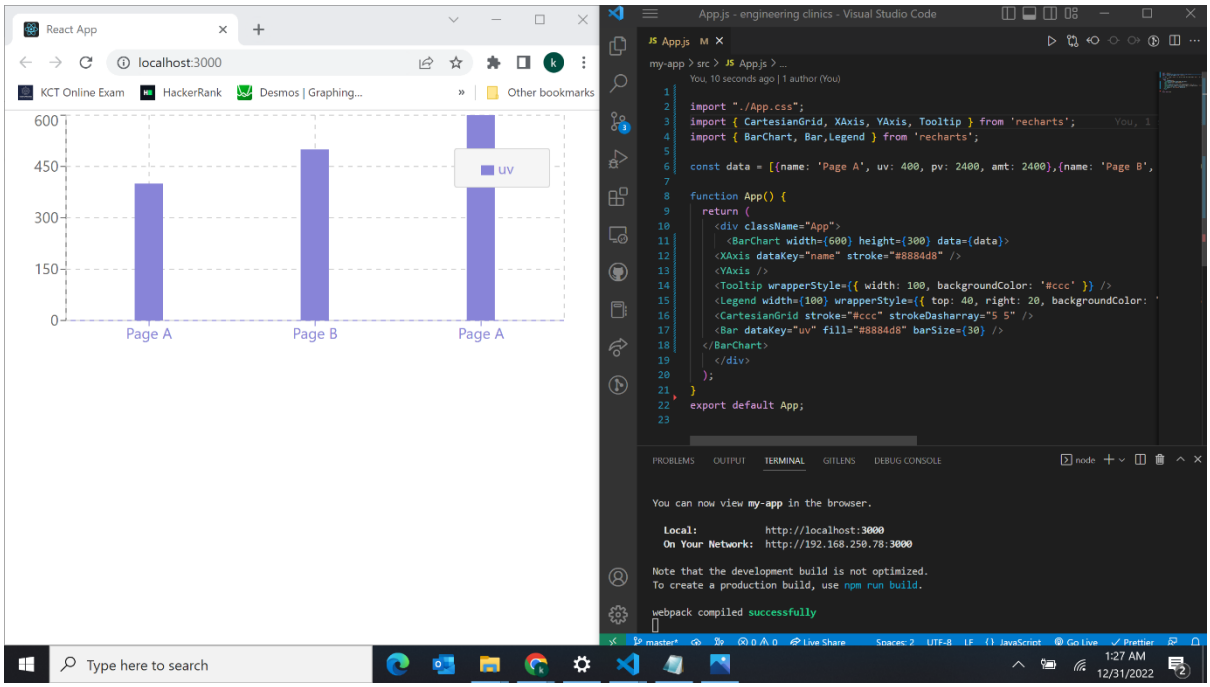
Visualization is the use of computer-supported, visual representation of data. Unlike static data visualization, interactive data visualization allows users to specify the format used in displaying data.

Common visualization techniques are

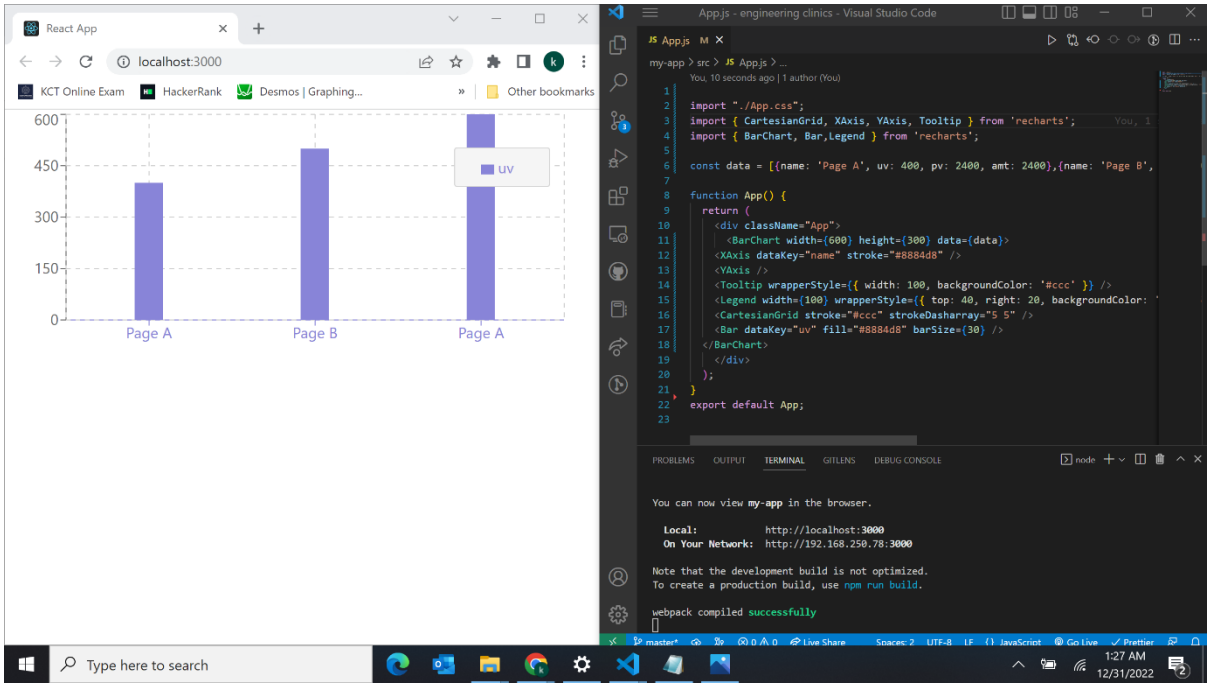
4.1.1: Line graph: This shows the relationship between items. It can be used to compare changes over a period of time.



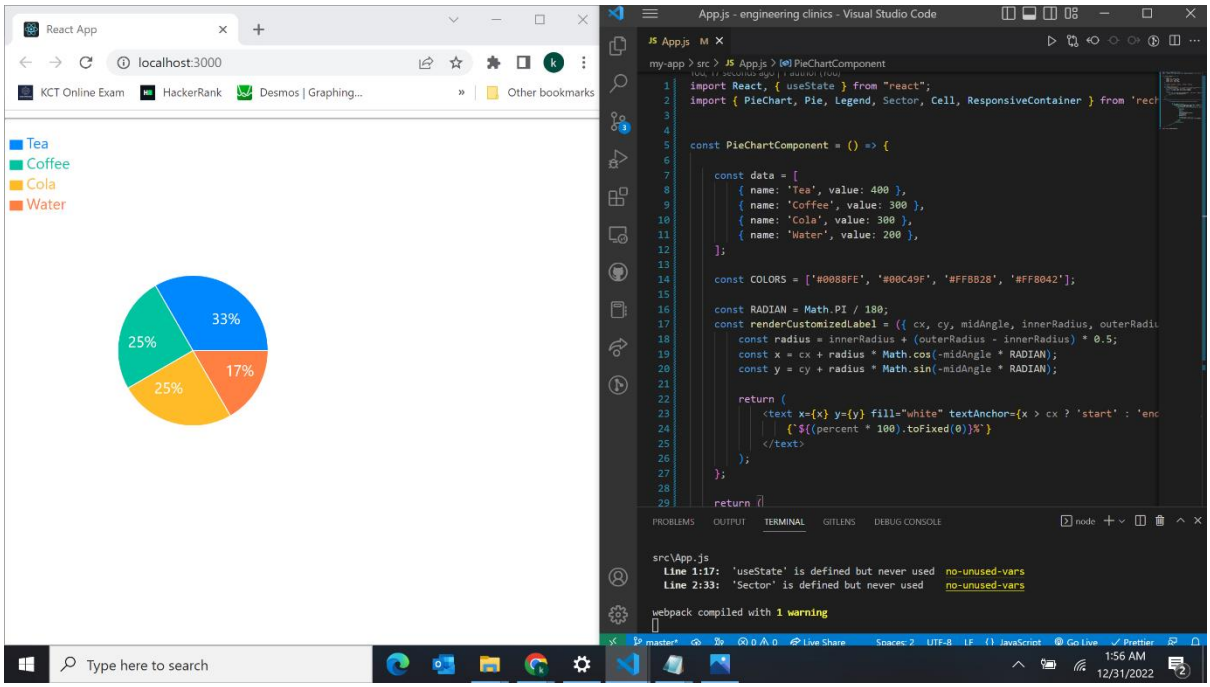
4.1.2: Bar chart: This is used to compare quantities of different categories.



4.1.3: Scatter plot: This is a two-dimensional plot showing variation of two items.



4.1.4: Pie chart: This is used to compare the parts of a whole.



Thus, the format of graphs and charts can take the form of bar chart, pie chart, line graph, etc. It is important to understand which chart or graph to use for your data. Data visualization uses computer graphics to show patterns, trends, and relationship among elements of the data. It can generate pie charts, bar charts, scatter plots, and other types of data graphs with simple pull-down menus and mouse clicks. Colors are carefully selected for certain types of visualization. When color is used to represent data, we must choose effective colors to differentiate between data elements.

In data visualization, data is abstracted and summarized. Spatial variables such as position, size, and shape represent key elements in the data. A visualization system should perform a data reduction, transform and project the original dataset on a screen.

4.2 METHODOLOGY

DATASET

For testing the framework, the dataset is taken from Kaggle. This project isn't dataset specific, it is visualization based, so any possible dataset will work fine.

DATA TRANSFORMATION

The selected dataset includes information that may or may not be helpful to us. Data pre-processing is the process of identifying the most important data from the whole dataset for subsequent processing. Hence, the unwanted part of the data is removed. As a result, processed data is

ready to train the machine.

FRONT-END API

It is the interface between the front-end application and the data visualization and machine learning model. It fetches the input provided by the user in the front end and throws it to visualization pipeline that works behind the scene to provide visual data, then carries out data analytics

BACK-END

The background operations are carried out in JavaScript which manages file upload, delete, visualizing and other user functions, assisted by Python for carrying out Data Analytics on the dataset.

JS-PYTHON CONNECTIVITY

JS takes data from the user and manages data modification to user preferences, The modified data is then feeded to Python running on server to carry out data analysis, The resulting model can the be accessed via JS.

CLIENT-SIDE SCRIPTING

It is designed in such a way to reduce the resource usage of client, data visualizing and other basic functionalities of the software that consume less resources are integrated with the client side, It also reduces the number of transactions between the client and server providing network optimality

SERVER-SIDE SCRIPTING

The software also utilizes the power of the server to provide quick results on training and Data Analytics and Machine Learning are carried out in the server to reduce the pressure on client-side resources, Delivering faster and efficient trained models within a short span of time regardless of the client's resources

DOCKERIZED ENVIRONMENT

It depends on various libraries of different versions for functioning, it is dockerized to provide easy to deploy and use functionality to clients. It also allows users to run parallel instances training multiple models parallely

CHAPTER 5

SYSTEM REQUIREMENTS

5.1 SOFTWARE REQUIREMENTS

- Vs Code
- Google Collab
- Git
- Docker

5.2 HARDWARE REQUIREMENTS

- Minimum 64 GB of free hard disk space
- Core i5/i7, AMD Ryzen 5/7
- RAM: 8GB

5.3 LIBRARIES & FRAMEWORKS

- ReactJS

5.4 PROGRAMMING LANGUAGES

- Python
- JS

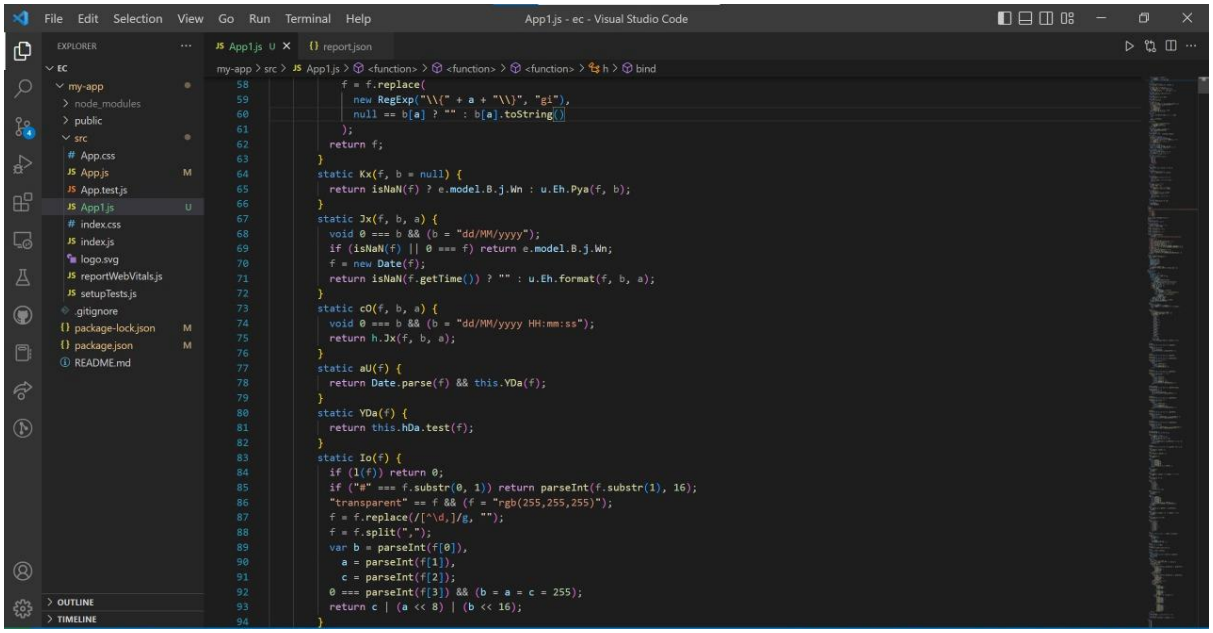
CHAPTER 6

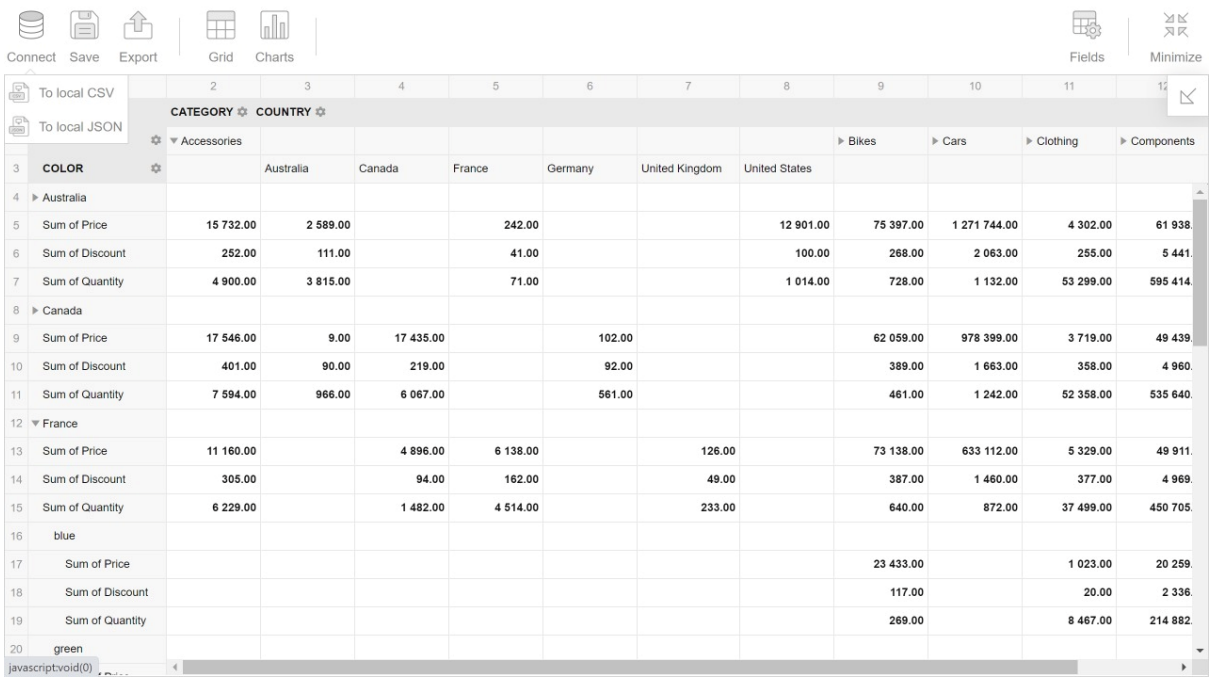
CONCLUSION

The expansion of businesses occurs at an exponential rate. Data analysis is essential for the promotion of such small firms. This approach of data visualisation is easier and simpler than sifting through the entire program's working code since it makes use of the user's privileges. Because it is necessary to download programmes from unauthorised and dubious third-party websites, average people lose their personal information and begin to question the system. Users can quickly and easily obtain accurate statistics and data visualisation using this strategy. This technique is also applicable to other situations where a significant volume of sensitive data needs to be analysed.

APPENDIX

CODE





Github : https://github.com/Dhanushkumar-S-G/No-Code-Analysis

REFERENCES

[1] Research Data Analysis with Power BI, Vijay Krishnan S Bharanidharan G Krishnamoorthy,

[2] The art of Data Analysis 1. Muhammad Ibrahim, Department of Statistics, Govt. MAO College, Lahore

[3] Characterizing Exploratory Visual Analysis: A Literature Review and Evaluation of Analytic Provenance in Tableau, Leilani Battle1 and Jeffrey Heer2,

[4] Towards Automated Data Cleaning Workflows, Ziawasch Abedjan Felix Neutatz Mohammad Mahdavi Larysa Visengeriyeva